Working papers series

WP ECON 06.19

Resolving Inconsistencies in Utility Measurement under Risk: Tests of Generalizations of Expected Utility

Han Bleichrodt (Erasmus University)

Jose Maria Abellan-Perpiñan (U. of Murcia)

Jose Luis Pinto Prades (U. Pablo de Olavide)

Ildefonso Mendez-Martinez (U. of Murcia)

JEL Classification numbers: I10.
Keywords: Utility Measurement, Nonexpected Utility, Prospect Theory, Health.

Department of Economics

Resolving Inconsistencies in Utility Measurement under Risk:
Tests of Generalizations of Expected Utility

Han Bleichrodt, Erasmus University, Rotterdam, The Netherlands.
Jose Maria Abellan-Perpiñan, University of Murcia, Spain.
Jose Luis Pinto-Prades, University Pablo de Olavide, Sevilla, Spain.
Ildefonso Mendez-Martinez, University of Murcia, Spain.

March 2006

Abstract

This paper explores inconsistencies that occur in utility measurement under risk when expected utility is assumed and the contribution that prospect theory and some other generalizations of expected utility can make to the resolution of these inconsistencies. We used five methods to measure utilities under risk and found clear violations of expected utility. Of the theories studied, prospect theory was **the** most consistent with our data. The main improvement of prospect theory over expected utility was in comparisons between a riskless and a risky prospect (riskless-risk methods). We observed no improvement over expected utility in comparisons between two risky prospects (risk-risk methods). An explanation for the latter observation may be that there was less distortion in probability weighting in the interval [0.10, 0.20] than has commonly been observed.

KEYWORDS: Utility Measurement, Nonexpected Utility, Prospect Theory, Health.

Address correspondence to: Han Bleichrodt, Dept. of Economics, H13-27, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. Email: bleichrodt@few.eur.nl.

Acknowledgements: We are grateful to Peter Wakker, an associate editor, and three reviewers for their helpful comments on previous drafts and to George Wu for sending us additional statistics on the data in Wu and Gonzalez (1996). Han Bleichrodt's research was made possible by a VIDI-grant from the Netherlands Organization for Scientific Research (NWO).

1. Introduction

Utility measurement is an important tool for decision analysis and helps clients to make better-informed choices. In order to obtain valid and consistent utility measurements, the decision analyst needs a theory that is descriptively valid so he can then use these utilities as inputs into a model, e.g. a decision tree, that will help the client to make a choice truly representing his preferences.

This paper explores inconsistencies in the measurement of utilities under risk and tries to find a theory of preferences that can solve these inconsistencies. Risky utilities are widely used in decision analysis. The main examples are the *probability equivalence* (PE) method, where people are asked to state a probability in a risky prospect that makes them indifferent between this risky prospect and a given outcome for sure, and the *certainty equivalence* (CE) method, where people are asked to state a certain outcome that makes them indifferent between this outcome and a given risky prospect. The appealing feature of using methods involving risk is that they allow the analyst to incorporate people's attitudes towards risk into decision analysis.

The common way to analyze the responses to utility measurements under risk is by assuming expected utility. A rationale for adopting expected utility is that decision analysis is essentially a prescriptive exercise and that expected utility is the dominant prescriptive theory of decision under risk. The problem with this point of view is that in most practical applications measuring utilities is a descriptive task, and the descriptive deficiencies of expected utility are widely documented (Starmer 2000). Using expected utility to analyze responses to utility measurement tasks in spite of its poor descriptive record carries the danger that biased utilities results and decision analyses based on these biased utilities result in incorrect recommendations.

Several studies have shown that utility measurement based on expected utility leads to inconsistencies. Hershey and Schoemaker (1985) showed that PE measurements result in systematically higher utilities than CE measurements. McCord and de Neufville (1986) found that the utility function elicited by the CE method depends on the value at which probability is fixed. In the health domain, several authors have shown that under expected utility assessment procedures that are theoretically equivalent

produce systematically different utilities (Llewellyn-Thomas et al. 1982, Rutten-van Mölken et al. 1995, Bleichrodt 2001, Oliver 2003, {Pinto-Prades, 2005 #657}). For example, Pinto-Prades and Abellan-Perpiñan (2005) found that under expected utility the utility of a given health state varied between 0.48 and 0.80 depending on the assessment procedure that was used. Such large discrepancies can have significant effects on the recommendations for practical decision analyses.

One solution for the observed inconsistencies was suggested by Bleichrodt et al. (2001). They argued that people's preferences are affected by bias (comment: bias rarely used as a plural noon). The types of bias they considered were (1) probability weighting, the nonlinear evaluation of probabilities, and (2) loss aversion, the finding that people are more sensitive to losses than to gains of the same size. Both have been modelled by prospect theory (Kahneman and Tversky 1979, Tversky and Kahneman 1992), currently the main descriptive theory of decision under risk. Bleichrodt et al. proposed new formulas based on prospect theory to evaluate answers to PE and CE measurements and showed that these new formulas were able to resolve the systematic discrepancy between PE and CE utilities.

The PE and CE are examples of methods in which a riskless prospect is compared with a risky prospect (*riskless-risk methods)*. Several empirical studies have observed that violations of expected utility primarily occur when one of the prospects under consideration is riskless; expected utility's descriptive record is much better when both prospects are risky (Conlisk 1989, Camerer 1992, Harless and Camerer 1994, Wu and Gonzalez 1996, Starmer 2000). Hence, it is of interest to examine the performance of prospect theory when we include methods that compare two risky prospects (*risk-risk methods*). It would be important for practical decision analysis in case we observed that expected utility could explain the data from risk-risk measurements. Then no information on probability weighting and loss aversion is required and preferences can be measured under expected utility by using risk-risk methods such as the lottery equivalence method (McCord and de Neufville 1986).

In this paper we compared five elicitation methods, three riskless-risk methods and two risk-risk methods. The aims of our study were threefold: first, to replicate the inconsistency between PE and CE under expected utility and to compare PE, and CE with a third riskless-risk method, second, to test

whether expected utility leads to consistent utilities for the two risk-risk methods, and, third, to explore whether prospect theory could explain any inconsistencies across the five methods that were observed under expected utility. Our study was motivated by the idea that if the five assessment procedures yield consistent results when analyzed under a particular preference theory, but inconsistent results under another, then the theory under which consistency is found is supported as a descriptive theory.

We studied the performance of expected utility and prospect theory in the health domain. The main reason we focused on health is practical relevance. Health is an important area of applied decision analysis (the majority of applied decision analyses have been in the health domain, see Keller and Kleinmuntz (1998) and Smith and von Winterfeldt (2004)) and risky methods are widely used in medical decision analysis.

Under expected utility, we found that the obtained utilities varied across the three riskless-risk methods and were larger than the obtained utilities under the risk-risk methods. Prospect theory fitted the data better than expected utility. The main improvement of prospect theory over expected utility was in the evaluation of the riskless-risk methods. For risk-risk methods we found no violations either for expected utility or for prospect theory and, hence, prospect theory did not improve over expected utility. Because neither expected utility nor prospect theory could entirely explain the data, we also considered several other generalizations of expected utility: rank-dependent utility (Quiggin 1981), disappointment aversion (Gul 1991), and two recently proposed gambling effect models (Bleichrodt and Schmidt 2002, Diecidue et al. 2004). However, none of these nonexpected utility models fitted the data as well as prospect theory.

In what follows, Section 2 introduces notation and briefly explains prospect theory. Section 3 describes the five elicitation methods used. Section 4 analyzes the predictions from expected utility and prospect theory and shows how utilities were computed under these two theories. Section 5 describes the experiment that we performed and Section 6 its results. Section 7 discusses the main findings and limitations of our study. Section 8 concludes the paper. The appendix contains derivations of results introduced in the text.

2. Notation and outline of the models

We study preferences over chronic health states. Chronic health states will be written as (Q,T), denoting T years in health state Q. Because the experiment reported later only invokes prospects with at most two different chronic health states, we will restrict the analysis to such *binary prospects*. Let $\left(p:(Q_1,T_1);\ (Q_2,T_2)\right)$ denote the prospect that gives (Q_1,T_1) with probability p and (Q_2,T_2) with probability 1−p. A prospect is *riskless* if p = 1, otherwise it is *risky*.

Preferences over prospects are denoted as usual: the relation $\succcurlyeq$ denotes weak preference, $\succ$ denotes strict preference, and ~ denotes indifference. Preferences over chronic health states correspond with preferences over riskless prospects. We assume throughout that prospects are *rank-ordered*, i.e. it is implicit in the notation $\left(p:(Q_1,T_1);\ (Q_2,T_2)\right)$ that $(Q_1,T_1) \succcurlyeq (Q_2,T_2)$. *Expected utility* holds if there exists a function U from the set of chronic health states to the real numbers, called the *utility function*, such that prospects $\left(p:(Q_1,T_1);\ (Q_2,T_2)\right)$ are evaluated by $pU(Q_1,T_1) + (1-p)U(Q_2,T_2)$ and preferences and choices correspond with this evaluation.

Prospect theory deviates in three important respects from expected utility. First, carriers of value are gains and losses relative to a *reference point*. The location of the reference point is exogenously given and is not specified by prospect theory. We will denote the reference point as (Q_0,T_0). In the formal analysis of Kahneman and Tversky (1979), where there is only one fixed reference point, the reference point is assigned utility zero. In this paper we will consider variations in the reference point and, hence, we do not follow this convention. A second deviation from expected utility is that people are more sensitive to losses than to corresponding gains, a phenomenon known as *loss aversion*. Third, people do not evaluate probabilities linearly, but weight probabilities; probability weighting for gains can be different from probability weighting for losses.

A prospect is *mixed* if it involves both a gain and a loss. A mixed prospect $(p:(Q_1,T_1); (Q_2,T_2))$, $(Q_1,T_1) \succ (Q_0,T_0) \succ (Q_2,T_2)$, $p \in (0,1)$, is evaluated as

$$PT(p:(Q_1,T_1); (Q_2,T_2)) = U(Q_0,T_0) + w^+(p)(U(Q_1,T_1) - U(Q_0,T_0)) \\ -\lambda\, w^-(1-p)(U(Q_0,T_0) - U(Q_2,T_2)), \quad (1)$$

where PT is the function assigning utility to a prospect under prospect theory assumptions, w^+ and w^- are probability weighting functions for gains and losses, and λ is a loss aversion parameter. The *probability weighting functions* assign weights of 0 and 1 to probabilities of 0 and 1, respectively, and are strictly increasing over this interval. We separate loss aversion from utility, because we consider varying reference points and we want to establish a link with expected utility. Therefore, the utility function U reflects the intrinsic utility of chronic health states. Our method for modelling loss aversion is similar to Shalev (2000) and Bleichrodt et al. (2001). In studies where the reference point is fixed, loss aversion is often incorporated in the utility function (Kahneman and Tversky 1979, Tversky and Kahneman 1992). In (1), outcomes are evaluated as deviations from the reference point through terms $U(Q_1,T_1) - U(Q_0,T_0)$ so as to combine the psychology of prospect theory with the utility function U of expected utility.

If $(Q_1,T_1) \succcurlyeq (Q_2,T_2) \succcurlyeq (Q_0,T_0)$, then

$$PT(p:(Q_1,T_1); (Q_2,T_2)) = U(Q_0,T_0) + w^+(p)(U(Q_1,T_1) - U(Q_0,T_0)) \\ + (1-w^+(p)(U(Q_2,T_2) - U(Q_0,T_0)). \quad (2)$$

If $(Q_0,T_0) \succcurlyeq (Q_1,T_1) \succcurlyeq (Q_2,T_2)$, then

$$PT(p:(Q_1,T_1); (Q_2,T_2)) = U(Q_0,T_0) - \lambda w^-(1-p)(U(Q_0,T_0) - U(Q_2,T_2))$$

$$- \lambda\big(1-w^{-}(1-p)\big)\big(U(Q_0,T_0) - U(Q_1,T_1)\big). \qquad (3)$$

Equation (3) is best interpreted as the dual of (2) with $1-w^{-}(1-p)$ instead of $w^{+}(p)$ and a loss aversion parameter λ added to all utility differences.

We assume throughout that changes in the reference point leave the probability weighting functions w^{+} and w^{-}, the utility function U, and the loss aversion parameter λ unchanged. Empirical studies have shown that the most common pattern for the probability weighting functions is inverse S-shaped, overweighting small probabilities and underweighting intermediate and large probabilities (Tversky and Kahneman 1992, Tversky and Fox 1995, Wu and Gonzalez 1996, Gonzalez and Wu 1999, Abdellaoui 2000, Bleichrodt and Pinto 2000). Tversky and Kahneman (1992) proposed the following one-parameter functional form for the probability weighting function:

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma}+(1-p)^{\gamma})^{1/\gamma}}, \qquad (4)$$

which has an inverse S-shape for γ between 0.27 and 1. Tversky and Kahneman (1992) found a median value for γ^{+} (the parameter for gains) of 0.61 and for γ^{-} (the parameter for losses) of 0.69. Later studies found comparable values for γ^{+} and γ^{-} both for monetary (Abdellaoui 2000) and for health outcomes (Bleichrodt and Pinto 2000). These values imply that the probability for which the probability weighting functions change from overweighting probabilities to underweighting probabilities, i.e. the probability for which $w(p) = p$, lies between 0.30 and 0.40. Tversky and Kahneman (1992) also estimated the loss aversion coefficient and found a median value for λ equal to 2.25. A comparable value was obtained by Bleichrodt et al. (2001).

3. Elicitation methods

We used five methods to measure the utility of health states, three riskless-risk methods and two risk-risk methods. Table 1 shows the elicitation methods that we used, the response that was elicited is printed in bold. The first three methods are the riskless-risk methods. The PE method elicited the probability p that made a subject indifferent between (Q,T) for certain and a risky prospect giving (FH,T) with probability p and Death with probability 1− p, where FH stands for full health. The CE method elicited the duration T_{ce} that made a subject indifferent between (Q,T_{ce}) for certain and a risky prospect giving (FH,T) with probability p and Death with probability 1− p. The *value equivalence* (VE) method elicited the duration T_{ve} that made a subject indifferent between (Q,T) for certain and a risky prospect giving (FH,T_{ve}) with probability p and Death with probability 1− p. To enhance comparability between the three riskless-risk methods, the selected value of the *gauge duration* T was the same in all riskless-risk methods and the response p elicited in the PE method was also used in the CE and the VE method.

Table 1: The five elicitation methods used

Method	Question
Riskless-Risk Methods	
Probability Equivalence (PE)	(Q,T) ~ (**p**:(FH,T); Death)
Certainty Equivalence (CE)	(Q,**T_{ce}**) ~ (p:(FH,T); Death)
Value Equivalence (VE)	(Q,T) ~ (p:(FH,**T_{ve}**); Death)
Risk-Risk Methods	
Probability Lottery Equivalence (PLE)	(0.35:(Q,T); Death) ~ (**r**:(FH,T); Death)
Value Lottery Equivalence (VLE)	(0.35:(Q,T); Death) ~ (0.35:(FH,**T_{vle}**); Death)

Note: FH stands for “full health”.

The final two methods are the risk-risk methods. The *probability lottery equivalence* (PLE) method elicited the probability r that made a subject indifferent between the risky prospect that gives (FH,T) with probability r and Death with probability 1− r and the risky prospect that gives (Q,T) with probability 0.35 and Death with probability 0.65. The *value lottery equivalence* (VLE) method elicited the duration T_{vle} that made a subject indifferent between the risky prospect that gives (FH,T_{vle}) with

probability 0.35 and Death with probability 0.65 and the risky prospect that gives (Q,T) with probability 0.35 and Death with probability 0.65. The gauge duration T was the same in the PLE and the VLE methods and was set equal to the value of T that was used in the riskless-risk methods. The use of probability 0.35, for which previous studies observed little probability distortion, in the risk-risk methods does not mean that probability weighting plays no role in these methods as r would generally be lower than 0.35 and would be affected by probability weighting according to previous empirical findings.

It is well known that indifference judgments tend to be affected by scale compatibility if the dimension on which indifference is established is changed. Scale compatibility predicts that the weight people assign to an attribute increases as this attribute is more compatible with the response scale used (Tversky et al. 1988, Delquié 1993, 1997). The effect of scale compatibility in the present experiment is ambiguous. For instance, in the PE method the response scale is probability and, consequently, scale compatibility predicts that people will focus on probability when evaluating prospects. However, people could either focus on the probability p of the good outcome of full health or on the probability 1–p of the bad outcome of Death and the direction of the bias due to scale compatibility depends on which probability people focus. We tried to control for compatibility effects by determining all indifferences through a choice-based task.

Another type of compatibility effect is *strategy compatibility*, according to which qualitative decision tasks, such as choice, are compatible with qualitative decision strategies, such as lexicographic ordering, and quantitative decision tasks, such as matching, are compatible with quantitative decision strategies (Fischer and Hawkins 1993). Prior research has shown that strategy compatibility is stronger than scale compatibility (Fischer and Hawkins 1993, Delquié 1997). Fischer et al. (1999) suggested that task-goals play a central role in the construction of preferences. Specifically, they proposed that the prominent attribute of an alternative is weighted more heavily in response tasks whose perceived goal is to differentiate between alternatives than in tasks whose perceived goal is to equate alternatives. Neither strategy compatibility nor the task-goal hypothesis predict a bias in our results, because we used the same decision task, the determination of indifferences through choices, in all five methods.

4. Predictions

We assumed throughout that people prefer a longer life duration to a shorter one, both in full health and in health state Q. Expected utility then predicts that, except for random error, we should observe that $T = T_{ce} = T_{ve}$.[1] The two risk-risk methods could not be directly compared with each other and with the riskless-risk methods because they involved different probabilities.

The predictions made by prospect theory depend on the location of the reference point. Hershey and Schoemaker (1985) and Bleichrodt et al (2001) conjectured that in a utility elicitation task in which a subject compares two prospects and has to create an equivalence by varying a probability or outcome level of one prospect, he will take as reference level another outcome of the prospect that remains constant in the equivalence judgment. Their conjecture has been corroborated by empirical evidence (Stalmeier and Bezembinder 1999, Morrison 2000, Bleichrodt et al., 2001, Robinson et al. 2001). In the PE and in the VE this argument implies that the reference point is (Q,T). In the CE the reference point is either (FH,T) or Death. The data in Bleichrodt et al. (2001) suggested that people take Death as their reference point in the CE questions.

Consequently, we would expect that $T = T_{ve}$ under prospect theory, because the reference point is the same in the PE and in the VE. However, T_{ce} may well differ from T. Under this hypothesis, the PE and VE method compare mixed prospects, whereas the CE compares either two prospects involving only gains or two prospects involving only losses. It is well-known that, due to loss aversion, people are more risk averse for mixed prospects than for prospects involving only gains or only losses (Payne et al. 1980, 1981) and, hence, we expect that $T_{ce} > T$. Again, because the risk-risk methods used different probabilities they do not yield specific predictions unless additional assumptions are made.

We obtained more conclusive tests of expected utility and prospect theory by computing health state utilities. To be able to compute health state utilities we assumed in all models *multiplicativity* of U: $U(Q,T) = H(Q)L(T)$, where H and L are real-valued utility functions over the set of health states and the set of life durations, respectively. Throughout, we used the scaling U(Death) = 0 and H(FH) =1.

Empirical support for multiplicativity was obtained by Miyamoto and Eraker (1988), Doctor et al. (2004), and Bleichrodt and Pinto (2005). A test of multiplicativity is obtained by comparing the answers to the PE questions for different gauge durations with each other and by comparing the answers to the PLE questions for different gauge durations with each other. Multiplicativity implies that, except for random response error, we should find the same probabilities p in the different PE questions and the same probabilities r in the different PLE questions.

To be able to compute H(Q) from the responses to the CE, VE, and VLE questions, we had to assume a specific form for L(T). We first assumed that L is linear in which case U(Q,T) is equal to the QALY model, the most widely used model in medical decision analysis. The first row of Table 2 shows the expression for H(Q) under expected utility with linear utility. There is little empirical support for the assumption that utility is linear in life duration; an exception is the study by Doctor et al. (2004). We, therefore, subsequently assumed that L is a power function. The power function is often used in decision analysis and several studies have observed that it yields a good fit in the health domain (e.g. Pliskin et al. 1980, Miyamoto and Eraker 1985, Stiggelbout et al. 1994, Cher et al. 1997). The second row of Table 2 shows the effect on H(Q) of replacing the assumption of linear utility by the assumption of power utility. The table shows that the PE and the PLE method are not affected by the choice of L.

The final row of Table 2 shows H(Q) under prospect theory with power utility for life duration. To be able to evaluate the five elicitation methods under prospect theory we had to specify the location of the reference point for each method. Here we followed the suggestions of Hershey and Schoemaker (1985) and Bleichrodt et al (2001) and took (Q,T) as the reference point in the PE and in the VE and either (FH,T) or Death in the CE. In the PLE we took either (FH,T), or (Q,T) or Death as the reference point. In the VLE, (FH,T_{vle}) was implausible as a reference point, because T_{vle} was varied to create an equivalence and we, therefore, only analyzed the data for reference points (Q,T) and death. To compute the probability weights, we assumed (6) with γ^+ the probability weighting parameter for gains and γ^- the probability weighting parameter for losses.

[1] This follows from transitivity and the assumption that more life duration is preferred to less.

Table 2: Utilities under expected utility (EU) and prospect theory (PT)

	PE	CE	VE	PLE	VLE
EU-linear	p	$p\frac{T}{T_{ce}}$	$p\frac{T_{ve}}{T}$	$\frac{r}{0.35}$	$\frac{T_{vle}}{T}$
EU-power	p	$p(\frac{T}{T_{ce}})^{\beta}$	$p(\frac{T_{ve}}{T})^{\beta}$	$\frac{r}{0.35}$	$(\frac{T_{vle}}{T})^{\beta}$
PT	A	*RP Death*: $w^{+}(p)(\frac{T}{T_{ce}})^{\beta}$	$A(\frac{T_{ve}}{T})^{\beta}$	*RP Death*: $\frac{w^{+}(r)}{w^{+}(0.35)}$	*RP Death*: $(\frac{T_{vle}}{T})^{\beta}$
		RP (FH,T): $(1-w^{-}(1-p))(\frac{T}{T_{ce}})^{\beta}$		*RP (Q,T)* B *RP (FH,T)* $\frac{1-w^{-}(1-r)}{1-w^{-}(0.65)}$	*RP (Q,T)* $(\frac{T_{vle}}{T})^{\beta}$

Note: RP stands for reference point. $A = \frac{w^{+}(p)}{w^{+}(p)+\lambda w^{-}(1-p)}$. $B = \frac{w^{+}(r)}{w^{+}(r)+\lambda(w^{-}(1-r)-w^{-}(0.65))}$

5. Experiment

Background

The subjects were sixty-five economics students (aged between 22 and 29) from the University of Murcia. They were paid €36 to participate in five experimental sessions, each lasting approximately one hour. In each experimental session a different method was elicited. The experiment was carried out in small group sessions with at most six subjects per group. The sessions were separated by at least one week. Prior to the actual experiment, the questionnaire was tested in several pilot sessions using university staff as subjects.

We elicited the utility of two health states. The health states were described through the EuroQol system, a widely used instrument to describe health states in medical research. The description of the health states is given in Table 3. Throughout the experiment, the health states were labelled as A and B.

Preferences were elicited through a sequence of choices. Empirical evidence suggests that determining indifferences through choices leads to fewer inconsistencies than determining indifferences through matching (Bostic et al. 1990), because choice-indifference methods tend to mitigate the amount of inconsistency due to changing the response dimension.[2]

Table 3: The description of health states A and B

Health state A	Health state B
• Some problems walking about • Some problems performing self-care activities (e.g. eating, washing, dressing) • No problems performing usual activities (e.g. work, study, family or leisure activities) • Moderate pain or discomfort • Moderately anxious or depressed	• Some problems walking about • Some problems performing self-care activities (e.g. eating, washing, dressing) • Unable to perform usual activities (e.g. work, study, family or leisure activities) • Moderate pain or discomfort • Moderately anxious or depressed

Details

Recruitment of subjects took place one week before the actual experiment started. At the recruitment, subjects received information about the experiment and were asked to read the descriptions of the two health states. In addition, the subjects were handed a practice question on the PE method. They were asked to answer this practice question at home. This procedure was intended to familiarize them with the PE method. Prior to the start of the first experimental session, during which the PE method was administered, the subjects were asked to explain their answer to the practice question. When we were not convinced that a subject understood the task, we explained it again until we were convinced that he

[2] It should be noted that our method for determining indifferences through choice differed from the PEST procedure, the psychometric estimation procedure that Bostic et al. used. It is, however, plausible that the advantages of using choice also accrue to the present study.

understood the task. The same procedure was repeated for each of the remaining experimental sessions. The subjects received a practice question to take home showing the method that would be administered in the next session, and before the actual experiment started they had to explain their answer to the question.[3] To motivate the subjects, we told them at the start of the first experimental session that their answers were important for health policy to determine priorities between medical treatments.

The order in which the methods were administered was: first session PE, second CE, third PLE, fourth VE, and fifth VLE. The experiment was part of a larger experiment. We assumed that the presence of the other experimental tasks and the delay of at least one week between the sessions made it unlikely that the subjects would recall their previous answers or would note the relationship between the sessions.

At the beginning of each experimental session, instructions were read aloud and an additional practice question was given. We asked six questions per method by combining each health state (A and B) with three values for the gauge duration T: 13, 24, and 38 years. We used life durations substantially lower than the subjects' life-expectancy to avoid perception problems: subjects may find it hard to perceive living for very long durations which exceed their life-expectancy. To avoid order effects, we varied the order in which the different questions were asked within a section. To minimize response errors, the subjects had to confirm the elicited indifference value after each question.

We also determined through a choice-based procedure the life duration T′ that made a subject indifferent between T years in health state A and T′ years in full health, for T equal to 13, 24, and 38 years. The same question was asked for health state B. These questions were included to test the appropriateness of assuming power utility for life duration. Under power utility, we should find that the ratio $\frac{T'}{T}$ is constant. Also, in the presence of multiplicativity, the condition that the ratio $\frac{T'}{T}$ is constant implies that the utility for life duration must be a power function (Doctor and Miyamoto 2003). These questions were asked in the second experimental session.

[3] An example of a PE question can be found in the electronic companion pages to this paper. The wording of the other questions was similar.

Analysis

We used the nonparametric Friedman test to test for significance of utility differences among the five methods and for significance of differences between T, T_{ce}, and T_{ve}. When the hypothesis of equality was rejected by the Friedman test, we performed multiple pairwise comparisons by the Wilcoxon signed-rank test. The Friedman test was also used to test for the appropriateness of assuming multiplicativity (p should be the same in the three PE questions and r should be the same in the three PLE questions) and the power function ($\frac{T'}{T}$ should be constant, see above). In all tests, we used a significance level of 1% to control for experimentwise Type I error, the phenomenon that when many tests on a given level are performed, some will be significant by chance (multiple significance testing).

A distribution-free algorithm was used to determine the optimal values of the parameters in expected utility and prospect theory for each subject separately based on the utilities elicited by the five methods. We started by setting each parameter equal to one, the case corresponding to expected utility with linear utility for life duration. Then we searched for the values of the parameters that minimized the sum of squared differences between the elicited utilities.[4] We varied β between 0.05 and 2, γ^+ and γ^- between 0.25 and 2, and λ between 0.25 and 4. Using wider bounds occasionally caused the program to choose extreme and implausible values so that all utilities were close to zero. The range of parameters used includes the estimates from the existing empirical literature. The optimal parameters were determined with an accuracy of 0.01.

[4] To examine the sensitivity of the results to outliers, we also determined the parameters that minimized the sum of absolute differences. There were only small differences between the two sets of estimates.

6. Results

Preliminaries

Two subjects were excluded from the analyses of health state A and 19 from the analyses of health state B because their choices implied that they did not always prefer more life-years to less. This left 63 and 46 subjects in the analyses of health states A and B, respectively. More of the subjects had to be excluded for health state B, because B is a worse health state than A. The worse a health state, the more likely there is a duration for which the subjects do not prefer additional life-years. The excluded subjects were those with the lowest utilities. The fact that more of the subjects were excluded for health state B than for health state A will not bias our conclusions, because these are not based on comparisons between the utilities for health states A and B.

Table 4: Median responses. Interquartile ranges in parentheses.

Health state A					
T	PE	CE	VE	PLE	VLE
13y.	0.58	17	17	0.16	6
	(0.54-0.63)	(14-22)	(11-21)	(0.12-0.19)	(4-6)
24y.	0.68	27	28	0.17	11
	(0.65-0.73)	(23-35)	(22-36)	(0.12-0.19)	(9-13)
38y.	0.72	42	46	0.17	18
	(0.68-0.77)	(38-48)	(41-48)	(0.13-0.18)	(16-19)

Health state B					
T	PE	CE	VE	PLE	VLE
13y.	0.49	18	14	0.13	4
	(0.45-0.55)	(13-25)	(10-18)	(0.11-0.16)	(3-6)
24y.	0.57	31	21	0.13	8
	(0.51-0.65)	(24-36)	(14-38)	(0.11-0.16)	(7-11)
38y.	0.59	44	43	0.13	14
	(0.50-0.68)	(39-48)	(32-51)	(0.11-0.14)	(12-17)

Table 4 shows the median responses and in parentheses the interquartile range of the responses. The mean responses were similar to the medians. Because health state A is better than health state B, we should observe higher responses in the PE, VE, PLE, and VLE methods and lower responses in the CE

methods for health state A. All the subjects satisfied this consistency requirement. Note that r was much lower than 0.35, and was within a range for which previous studies found overweighting of probabilities. Hence, probability weighting could affect the PLE questions.

Contrary to the predictions of expected utility, we could reject equality of T, T_{ce}, and T_{ve} for both health states and for all three gauge durations ($P < 0.01$). The finding that T_{ce} significantly exceeded T ($P < 0.001$ in all cases) is consistent under prospect theory with more risk aversion for mixed prospects. Contrary to prospect theory, we also found that T_{ve} generally exceeded T with the exception for health state B with the gauge duration T equal to 24 years. The difference was always significant for health state A ($P < 0.001$). For health state B the difference was marginally significant for T = 13 years and T = 38 years ($P = 0.035$ and $P = 0.025$) and not significant for T = 24 years ($P = 0.650$).

The tests of multiplicativity yielded mixed results. For both health states we could reject the hypothesis that the probabilities in the three PE questions were equal ($P < 0.01$). However, we could not reject the hypothesis that the probabilities in the three PLE questions were equal ($P = 0.148$ for health state A and $P = 0.085$ for health state B). The tests of the appropriateness of using power utility for life duration yielded positive results: for both health states we could not reject the hypothesis that the ratio $\frac{T'}{T}$ was constant ($P = 0.565$ for health state A, $P = 0.085$ for health state B).

Main findings

Figure 1 shows the median utilities under expected utility with linear utility for life duration. For both health states and for all gauge durations we found significant differences between the five methods ($P < 0.001$ in all cases). The difference between the utilities was generally considerable with a maximum value of 0.39 (VE−VLE for health state A and gauge duration 38 years). The typical pattern was VE > PE > CE > PLE > VLE. The differences between PLE and VLE were, however, not significant. With few exceptions, all other paired differences were significant. Hence, we found that riskless-risk methods

yielded inconsistent results, that riskless-risk methods led to higher utilities than risk-risk methods, and that risk-risk methods led to consistent results under expected utility with linear utility for life duration.

The first row of Table 5 shows the medians of the individual estimates for the power function parameter under expected utility for each health state and for each gauge duration separately. The estimates reflect a substantial degree of concavity of the utility function for life duration. We could clearly reject the hypothesis that the five methods yield the same utilities under expected utility with the optimal power coefficients ($P < 0.001$ for both health states and for all gauge durations). The consistency between PLE and VLE that we observed under expected utility with linear utility for life duration no longer holds: all differences between PLE and VLE were significant under expected utility with power utility for life duration. Hence, the systematic differences between the five methods could not be explained by the assumed linearity of the utility for life duration only.

Figure 1: Median utilities under expected utility with linear utility for life duration

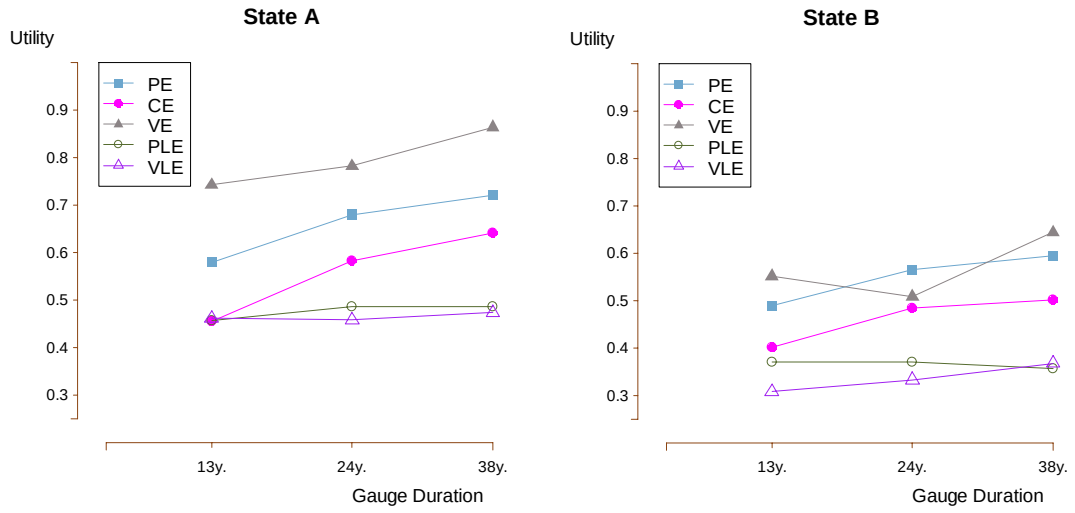


Table 5 also displays the medians of the individual parameter estimates for prospect theory where the reference point in the CE, PLE, and VLE methods is Death. We also analyzed the data under the other possible reference points, but these did not lead to smaller differences between the methods and the convergence and fit at the individual level was far worse. Compared with other studies, we observed less

distortion in probability weighting and less loss aversion. The finding of less probability weighting is not consistent with Rottenstreich and Hsee (2001), who found that the probability weighting function is more curved for "affect-rich" outcomes, such as health, compared to "affect poor" outcomes like money. Utility for life duration is concave; the estimates for the power coefficient agree with the findings from previous studies on the utility for life duration under nonexpected utility. There was considerable variation in the estimates at the individual level. The mean lengths of the interquartile range were 0.498, 1.509, 0.322, and 0.400 for β, λ, γ^+, and γ^-, respectively.

Table 5: Medians of the individual parameter estimates

	Health state A			**Health state B**		
Model / **Duration**	**13**	**24**	**38**	**13**	**24**	**38**
EU-power model						
β	0.46	0.45	0.48	0.60	0.53	0.60
Prospect Theory						
β	0.86	0.73	0.65	0.78	0.65	0.65
λ	2.13	2.00	1.84	1.83	1.77	1.53
γ^+	0.93	0.79	0.73	0.95	0.84	0.77
γ^-	0.80	0.80	0.80	0.90	1.20	0.80

We further examined the data under prospect theory with the probability weighting and loss aversion parameters obtained by Tversky and Kahneman (1992): $\gamma^+ = 0.61$, $\gamma^- = 0.69$, and $\lambda = 2.25$. We analyzed this case because Bleichrodt et al. (2001) were able to remove all discrepancies between PE and CE utilities using Tversky and Kahneman's values.

Neither of two prospect theories that we examined could fully explain the data. The differences between the five methods were significant for both versions of prospect theory ($P < 0.01$) except for health state B under prospect theory with the estimated optimal parameters where the differences between the methods were only marginally significant ($P = 0.040$, $P = 0.036$, and $P = 0.034$ for gauge durations 13 years, 24 years, and 38 years, respectively).

Table 6 shows the number of significant pairwise differences between the five elicitation

methods.[5] For each gauge duration, there were 10 comparisons between methods (PE versus CE, PE versus VE, etc.). The table shows that there were many inconsistencies under expected utility with linear utility for life duration. Using power utility instead of linear utility did not improve the performance of expected utility. Prospect theory with the probability weighting and loss aversion parameters obtained by Tversky and Kahneman (1992) performed better, even though many differences remained significant. [6] The number of significant differences was lowest under prospect theory with the optimal parameters. The performance of prospect theory with the optimal parameters was particularly good for health state B. Of course, in interpreting these results one should keep in mind that prospect theory with the optimal parameters had a greater degree of freedom than the other theories.

Table 6: Number of significant pairwise differences between methods based on median parameters and a significance level of 1%

	Health state A			Health state B		
Model / Duration	**13**	**24**	**38**	**13**	**24**	**38**
EU-linear	7	7	9	6	7	8
EU-power model	9	8	9	6	6	7
Prospect Theory TK	6	7	6	4	4	6
Prospect Theory Opt.	3	4	4	1	2	1

Note: Prospect Theory TK stands for Prospect Theory with the probability weighting and loss aversion parameters obtained by Tversky and Kahneman (1992)

Figure 2 shows the results under prospect theory with Tversky and Kahneman's (1992) values and linear utility for life duration. The figure shows that the main problem is that PLE was too high. As can be seen from Table 2, this problem cannot be solved by allowing for utility curvature because neither the PE nor the PLE are affected by the assumed form of the utility for life duration and, hence, their discrepancy will remain after correction for utility curvature. Under prospect theory, the difference between the PLE and the other methods can only be explained by a difference in the degree of probability weighting in our

[5] Tables with all the P-values from the Wilcoxon tests are in the electronic companion pages.
[6] Data in the Table are under linear utility for life duration. Similar data was obtained when we allowed for curved utility for life duration.

study as compared with Tversky and Kahneman (1992). Recall from Table 2 that the utility according to the PLE was equal to $\frac{w^+(r)}{w^+(0.35)}$. Because the PLE utilities were too high compared with the utilities elicited through the other four methods, this ratio was too high when Tversky and Kahneman's parameter values were used. Given that the response r in the PLE was generally between 0.10 and 0.20, our data suggest less overweighting of probabilities in that range than suggested by Tversky and Kahneman (1992) assuming that there is comparable (absence of) probability weighting around 0.35.

Figure 2: Median utilities under prospect theory with Tversky and Kahneman's values and linear utility for life duration

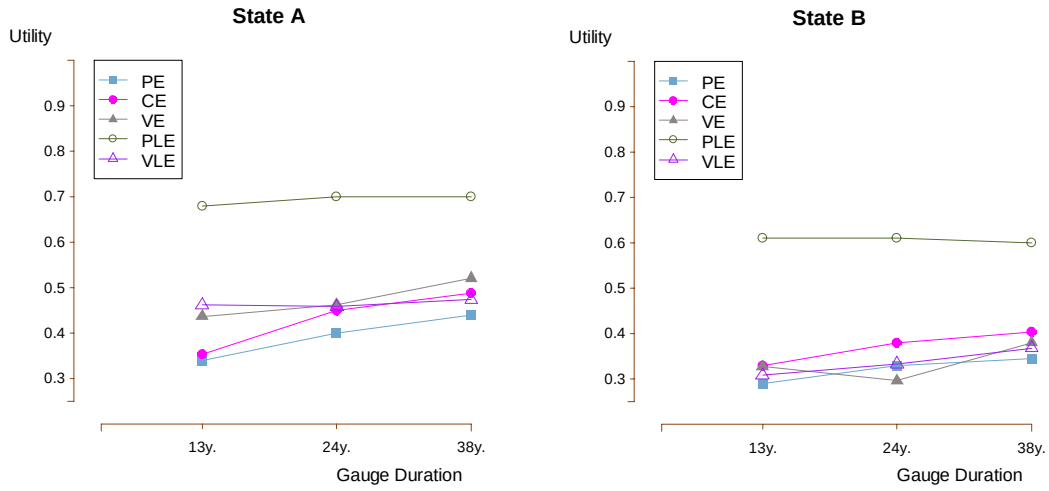
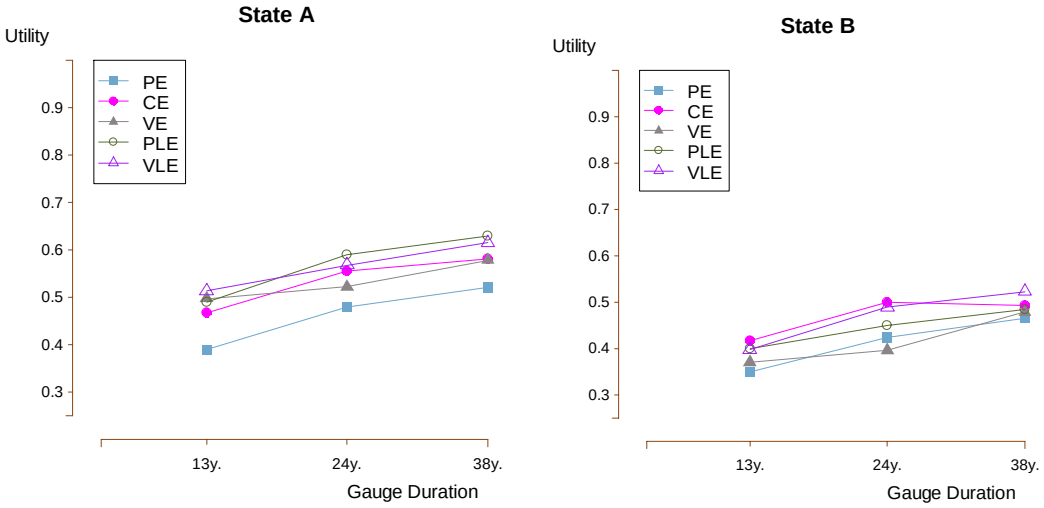


Figure 2 also shows that prospect theory could rather well explain the differences between the PE, CE, VE, and VLE except that for health state A the PE was a bit too low. In these four methods, the probabilities involved generally exceeded 0.35, suggesting comparable probability weighting in the range [0.35,1] as found by Tversky and Kahneman (1992). The finding that CE generally exceeded PE is in contrast with Bleichrodt et al. (2001). The difference between PE and CE was significant for health state A and gauge durations 24 years and 38 years and for health state B and gauge duration 38 years.

Figure 3 shows the median utilities under prospect theory with the optimal parameters. In general the utilities were close; the exception is that for health state A the PE was too low. For health state B there seem to be no systematic differences between the five methods.

Figure 3: Median utilities under prospect theory with optimal parameters



Finally, we examined the fit of the various theories by imposing on each individual the median optimal parameters for each health state-gauge duration pair and by then examining which theory yielded the lowest sum of squared errors between the five assessment methods. Table 7 reports the proportion of subjects for whom each model provided a superior fit to the data. The table shows that for most subjects prospect theory with the optimal parameters was most consistent with their data. These findings are not distorted by differences in degrees of freedom between the models because we imposed the median preferences on each subject.[7]

[7]There is a small caveat here, Because the median optimal parameters varied across health state-gauge duration pairs, expected utility with power utility and prospect theory with the optimal parameters had some additional flexibility.

Table 7: Proportion of individuals for whom a particular model fitted best in terms of the sum of squared residuals based on the median parameter estimates

Model Health state	EU linear	EU power	PT TK	PT opt
A, 13 years	12.7	38.1	7.9	41.3
A, 24 years	12.7	19.0	15.9	52.4
A, 38 years	7.9	9.5	12.7	69.8
B, 13 years	10.9	28.3	2.2	58.7
B, 24 years	4.3	26.1	10.9	58.7
B, 38 years	0	13.0	6.5	80.5

Auxiliary analyses

Although prospect theory was more consistent with the the data than expected utility, it could not entirely explain them and, hence, we also examined the data under four other nonexpected utility models. The models we considered were rank-dependent utility (Quiggin 1981), Gul's (1991) theory of disappointment aversion, and two recently proposed gambling effect models (Bleichrodt and Schmidt 2002, Diecidue et al. 2004). Rank-dependent utility is the special case of prospect theory where there is no loss aversion, i.e. Equation (2). Disappointment aversion is the special case of rank-dependent utility where the probability weighting function is equal to $\frac{p}{1+(1-p)\delta}$. The parameter $\delta \in (-1,\infty)$ reflects disappointment aversion. The two gambling effect models deviate from expected utility by assuming that there is not one utility function over outcomes but two. In the model of Diecidue et al. (2004) preferences are *prospect-dependent*: if a prospect is risky its outcomes are evaluated by a utility function U, if it is riskless its outcomes are evaluated by a utility function V. In the model of Bleichrodt and Schmidt (2002) preferences are *context-dependent*: when both prospects in a comparison are risky their outcomes are evaluated by a utility function U, otherwise the outcomes of both prospects are evaluated by a utility function V. An interesting property of the gambling effect models is that they predict that under expected utility riskless-risk methods lead to higher utilities than risk-risk methods and that expected utility will give consistent results in risk-risk methods, two predictions that were confirmed by the data.

None of the four models fitted the data as well as prospect theory with the optimal parameters. The differences between the five methods were significant for all theories ($P < 0.01$ in all cases), except for health state B and gauge duration 24 years in the model of Bleichrodt and Schmidt ($P = 0.084$). The number of significant pairwise differences between the five methods was in all models higher than under prospect theory with the optimal parameters, although this number was also low for health state B under Bleichrodt and Schmidt's (2002) gambling effect model. Finally, prospect theory with the optimal parameters was clearly the best fitting model when we imposed the median optimal estimates on each subject and then examined for each subject which model provided the superior fit. Details on the operationalization, the formulas for H(Q), the parameter estimates, the results from the auxiliary analyses and figures for rank-dependent utility, disappointment aversion, and the two gambling effect models are in the electronic companion pages.

7. Discussion

Our findings confirm that methods that are equivalent according to expected utility, produce systematically different results. The data suggests that evaluating riskless-risk methods through expected utility leads to utilities that are too high. We found no significant differences between risk-risk methods under expected utility (when the utility for duration is linear) and our data seems to add to the evidence that violations of expected utility primarily occur when one of the prospects under evaluation is riskless. Of the nonexpected utility models we studied, prospect theory with parameters tailored to the specific sample was most consistent with the data. Prospect theory could explain the systematic discrepancies between the riskless-risk methods. The finding that the two risk-risk methods yielded comparable results under expected utility is, however, harder to reconcile with prospect theory. One explanation for this finding may be that there was little overweighting of probabilities in the interval [0.10, 0.20].

It may be too optimistic to assume that one single model could explain all data. After all, any theory is necessarily incomplete and restricted in its scope (Payne et al. 1999). The number of deviations from expected utility that we considered was limited and other forms of bias will most likely have affected

people's responses. In particular, we ignored the impact of scale compatibility. Scale compatibility could explain, for instance, why T_{ve}, the response to the VE question, generally exceeded T, a finding that none of the theories considered could explain.

To operationalize prospect theory, we had to make assumptions about the location of the reference point. Empirical evidence seems to support our assumptions about the location of the reference point in the riskless-risk methods. No evidence exists about the location of the reference point in the risk-risk questions. Here we extended the arguments of Hershey and Schoemaker (1985) and Bleichrodt et al. (2001). The formation of a reference point may, however, be more complicated when both prospects are risky. For example, people may take a risky prospect as their reference point rather than a single outcome. It might be that in the PLE method, where the probability r was determined that made a subject indifferent between the risky prospects (r:(FH,T); Death)) and (0.35:(Q,T); Death)), the reference point for the best outcome of the prospects was (Q,T), but for the worst outcomes it was Death. It is not clear how to cover such a situation. In particular, it is not clear how to model probability weighting when the reference point is a risky prospect. Sugden (2003) presented such an extension for the case where people do not weight probabilities.

Our study was motivated by the idea that if five measurement methods yield consistent results under a particular preference theory then this theory is supported as a descriptive theory of decision under risk. There does not exist a gold standard for utility, however, and reconciliation between different measurements of utility suggests but does not prove that a particular preference theory is closer to people's true preferences. Whether other ways of assessing the descriptive validity of preference theories, e.g. by asking direct isolated choices, will produce similar conclusions is obviously an open question.

A drawback of using health outcomes is that we had to assume multiplicativity to be able to compute utilities. We observed mixed evidence on multiplicativity. It should be noted that even when multiplicativity was violated we could still compare the PE utilities with the PLE utilities. In that case we compared the utilities U(Q,T) under PE and PLE rather than the utilities H(Q). Our conclusions were not

affected when we only compared the PE with the PLE. Thus, the assumption of multiplicativity does not seem to be critical in the findings of our study.

Because we elicited preferences over health, the outcomes in our study had to be hypothetical. Several studies have addressed the question whether response patterns differ between questions with hypothetical outcomes and questions with real outcomes; see Camerer and Hogarth (1999) and Hertwig and Ortmann (2001) for extensive reviews. These studies used moderate monetary amounts as outcomes. The general conclusion from these studies is that the effect of real incentives varies across decision tasks. For the kind of tasks that we asked our subjects to perform, the determination of indifferences between binary prospects through choices, there appears to be no systematic difference in the general pattern of responses, although real incentives tend to reduce data variability.

Another drawback of using health outcomes could be that subjects had problems imagining the health states. We used the most common way to describe health states in medical research, but these descriptions are admittedly abstract and may have caused problems of imagination and, consequently, unreliable answers.

The use of students as subjects may limit the generalizability of our findings. Empirical evidence on health utility measurement has suggested, however, that there are no systematic differences in the patterns of responses obtained using convenience samples and those obtained using representative samples from the general population. For a review see de Wit et al. (2000).

A limitation of our study is that we used a fixed order in which the five methods were administered. This may have affected the results and it would have been better to change the order in which the methods were administered. We could exclude, however, the possibility that people behaved more in agreement with expected utility in later sessions, contradicting the hypothesis that more experience may lead to fewer violations of expected utility.

8. Implications

Let us finally discuss the implications of our findings for decision analysis practice. Our findings corroborate and extend earlier findings that different assessment procedures yield inconsistent results under expected utility. The differences are generally substantial and can be expected to affect the outcomes of practical decision analyses. In particular, our findings suggest that the common practice in (medical) decision analysis to measure utilities under risk by probability equivalence or certainty equivalence methods and to evaluate the responses through expected utility will lead to utilities that are biased upwards and, hence, to biased recommendations. The best way to solve these inconsistencies is to adopt a constructive preference approach and to solve the inconsistencies in an interactive process. Often such an approach is not possible, however. For example, in medical decision analysis utilities are commonly measured by medical staff who lack the time and training to solve the inconsistencies in utility measurement. Then the results of this paper may be useful.

At the aggregate level, our results show that expected utility should not be used to evaluate riskless-risk methods because the resulting utilities will be too high. Instead, these methods should be evaluated by prospect theory. The parameters found by Tversky and Kahneman (1992) performed rather well for riskless-risk methods although our data suggest somewhat less distortion in probability weighting and less loss aversion than implied by Tversky and Kahneman's parameters. For risk-risk methods we found no inconsistencies under expected utility. The absence of inconsistencies under expected utility can be explained by the virtual absence of overweighting of probabilities in the interval [0.10, 0.20], however, and we therefore feel that we cannot recommend using expected utility to evaluate risk-risk methods on the basis of our findings alone.

At the individual level the picture is more complex, because we found considerable variation in optimal parameter estimates. Here the best strategy seems to use several assessment methods simultaneously. Our data can help in the selection of these methods. For example, in medical decision analysis the PE and the PLE may be good choices because these require no assumptions about utility for life duration. Alternatively, the CE and the PLE may be selected because our data suggest that these are

not affected by loss aversion. Finally an argument in favour of using the VE and the VLE can be that these methods yielded the largest differences in elicited utilities under expected utility. If the selected assessment procedures give different results then the first step is to verify whether these differences affect the recommendations of the decision analysis. If so, and assuming that the constructive preference approach is not feasible, then the results of this paper can be useful in determining which theory is most consistent with a client's responses. The paper has derived and displayed patterns between assessment procedures and checking these patterns will improve the representation of the client's true preferences. If the patterns agree with the patterns observed in this paper then our individual results (Table 7) suggest that it is better to evaluate the client's responses by prospect theory with the parameters that we obtained than by expected utility even though this implies imposing the same preferences on all clients and ignoring the substantial variation in individual estimates.

Several concerns can be raised about the above recommendations, which we discussed in the previous section. These concerns are important and should be addressed in future research. However, for the progression of the field it is equally important to propose alternatives and advancements. We believe that the suggestions made above will lead to improvement of the common practice in (medical) decision analysis of using riskless-risk methods and analyzing these by expected utility.

Appendix: Derivation of the formulas in Tables 2 and 8.

Throughout we assume multiplicativity and power utility for life duration $U(Q,T) = H(Q)T^{\beta}$ and we use the scaling $H(FH) = 1$ and $U(Death) = 0$. Expected utility with linear utility is the special case of expected utility with power utility where $\beta = 1$.

PE method

The indifference (Q,T) ~ (p:(FH,T); Death) yields under expected utility $H(Q)T^{\beta} = pT^{\beta}$ or $H(Q) = p$. Under prospect theory, (1) yields $H(Q)T^{\beta} = H(Q)\ T^{\beta} + w^{+}(p)(T^{\beta} - H(Q)T^{\beta}) - \lambda w^{-}(1-p)\ H(Q)T^{\beta}$. Rearranging and deleting the common term T^{β} gives $H(Q) = \frac{w^{+}(p)}{w^{+}(p)+\lambda w^{-}(1-p)}$.

CE method

The indifference $(Q,T_{ce}) \sim (p:(FH,T);\ Death)$ yields under expected utility $H(Q)T_{ce}^{\beta} = pT^{\beta}$ and thus $H(Q) = p(\frac{T}{T_{ce}})^{\beta}$. Under prospect theory with reference point Death, $H(Q)T_{ce}^{\beta} = w^{+}(p)T^{\beta}$ and thus $H(Q) = w^{+}(p)(\frac{T}{T_{ce}})^{\beta}$. If the reference point is (FH,T), (3) gives $T^{\beta} - \lambda(H(Q)T_{ce}^{\beta} - T^{\beta}) = T^{\beta} - \lambda w^{-}(1-p)(-T^{\beta})$. Deleting common terms and rearranging gives $H(Q) = (1-w^{-}(1-p))(\frac{T}{T_{ce}})^{\beta}$.

VE method

The indifference $(Q,T) \sim (p:(FH,T_{ve});\ Death)$ yields under expected utility $H(Q)T^{\beta} = p\ T_{ve}^{\beta}$ and thus $H(Q) = p(\frac{T_{ve}}{T})^{\beta}$. Prospect theory with reference point (Q,T) gives by (1), $H(Q)T^{\beta} = H(Q)\ T^{\beta} + w^{+}(p)(^{\beta}_{ve} - H(Q)T^{\beta}) - \lambda w^{-}(1-p)\ H(Q)T^{\beta}$. Rearranging gives $H(Q) = \frac{w^{+}(p)}{w^{+}(p)+\lambda w^{-}(1-p)}(\frac{T_{ve}}{T})^{\beta}$.

PLE method

The indifference $(0.35:(Q,T);\ Death) \sim (r:(FH,T);\ Death)$ yields under expected utility $0.35H(Q)T^{\beta} = rT^{\beta}$, or $H(Q) = \frac{r}{0.35}$. Under prospect theory with reference point Death $w^{+}(0.35)H(Q)T^{\beta} = w^{+}(r)T^{\beta}$, or $H(Q) = \frac{w^{+}(r)}{w^{+}(0.35)}$. If the reference point is (Q,T), (1) and (3) give $H(Q)T^{\beta} - \lambda w^{-}(0.65)\ H(Q)T^{\beta} = H(Q)T^{\beta} + w^{+}(r)(T^{\beta} - H(Q)T^{\beta}) - \lambda w^{-}(1-r)\ H(Q)T^{\beta})$. Rearranging gives $H(Q) = \frac{w^{+}(r)}{w^{+}(r)+\lambda(w^{-}(1-r)-w^{-}(0.65))}$. If the

reference point is (FH,T), (3) gives $T^{\beta} - \lambda w^{-}(1-r)T^{\beta} = T^{\beta} - \lambda w^{-}(0.65)T^{\beta} - \lambda(1-w^{-}(0.65)(T^{\beta}-H(Q)T^{\beta})$. Rearranging gives $H(Q) = \frac{1-w^{-}(1-r)}{1-w^{-}(0.65)}$.

VLE method

The indifference (0.35:(Q,T); Death) ~ (0.35:(FH,T_{vle}); Death) yields under expected utility $0.35H(Q)T^{\beta} = 0.35\ T_{vle}^{\beta}$, or $H(Q) = (\frac{T_{vle}}{T})^{\beta}$. Prospect theory with reference point Death gives $w^{+}(0.35)H(Q)T^{\beta} = w^{+}(0.35)T_{vle}^{\beta}$ or $H(Q) = (\frac{T_{vle}}{T})^{\beta}$. If the reference point is (Q,T), (1) and (3) give $H(Q)T^{\beta} - \lambda w^{-}(0.65)\ H(Q)T^{\beta} = H(Q)T^{\beta} + w^{+}(0.35)(\ T_{vle}^{\beta} - H(Q)T^{\beta}) - \lambda w^{-}(0.65)\ H(Q)T^{\beta})$. Deleting common terms and rearranging gives $H(Q) = (\frac{T_{vle}}{T})^{\beta}$. If the reference point is (FH,T) then (3) gives $T_{vle}^{\beta} - \lambda w^{-}(0.65)T_{vle}^{\beta} = T_{vle}^{\beta} - \lambda w^{-}(0.65)T_{vle}^{\beta} - \lambda(1-w^{-}(0.65)(T_{vle}^{\beta}-H(Q)T^{\beta})$. Deleting common terms and rearranging gives $H(Q) = (\frac{T_{vle}}{T})^{\beta}$.

References

Abdellaoui, M. 2000. Parameter-free elicitation of utilities and probability weighting functions. *Management Science* **46** 1497-1512.

Bleichrodt, H. 2001. Probability weighting in choice under risk: An empirical test. *Journal of Risk and Uncertainty* **23** 185-198.

Bleichrodt, H., J. L. Pinto. 2000. A parameter-free elicitation of the probability weighting function in medical decision analysis. *Management Science* **46** 1485-1496.

Bleichrodt, H., J. L. Pinto. 2005. The validity of qalys under non-expected utility. *The Economic Journal* **115** 533-550.

Bleichrodt, H., J. L. Pinto, P. P. Wakker. 2001. Using descriptive findings of prospect theory to improve the prescriptive use of expected utility. *Management Science* **47** 1498-1514.

Bleichrodt, H., U. Schmidt. 2002. A context-dependent model of the gambling effect. *Management Science* **48** 802-812.

Bostic, R., R. J. Herrnstein, R. D. Luce. 1990. The effect on the preference reversal of using choice indifferences. *Journal of Economic Behavior and Organization* **13** 193-212.

Camerer, C. 1992. Recent tests of generalizations of expected utility theory. W. Edwards, eds. *Utility: Theories, measurement and applications*. Kluwer Academic Publishers, Boston, M.A., 207-251.

Camerer, C. F., R. M. Hogarth. 1999. The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty* **19** 7-42.

Cher, D. J., J. Miyamoto, L. A. Lenert. 1997. Incorporating risk attitude into markov-process decision models. *Medical Decision Making* **17** 340-350.

Conlisk, J. 1989. Three variants on the allais paradox. *American Economic Review* **79** 392-407.

de Wit, G. A., J. J. van Busschbach, F. T. de Charro. 2000. Sensitivity and perspective in the valuation of health status. *Health Economics* **9** 109-126.

Delquié, P. 1993. Inconsistent trade-offs between attributes: New evidence in preference assessment biases. *Management Science* **39** 1382-1395.

Delquié, P. 1997. 'bi-matching': A new preference assessment method to reduce compatibility effects. *Management Science* **43** 640-658.

Diecidue, E., U. Schmidt, P. P. Wakker. 2004. The utility of gambling reconsidered. *Journal of Risk and Uncertainty* **29** 241-259.

Doctor, J. N., H. Bleichrodt, J. Miyamoto, N. R. Temkin, S. Dikmen. 2004. A new and more robust test of qalys. *Journal of Health Economics* **23** 353-367.

Doctor, J. N., J. Miyamoto. 2003. Deriving quality-adjusted life-years (qalys) from constant proportional time tradeoff and risk posture conditions. *Journal of Mathematical Psychology* **47** 557-567.

Fischer, G. W., Z. Carmon, D. Ariely, G. Zauberman. 1999. Goal-based construction of preferences: Task goals and the prominence effect. *Management Science* **45** 1057-1075.

Fischer, G. W., S. A. Hawkins. 1993. Strategy compatibility, scale compatibility, and the prominence effect. *Journal of Experimental Psychology: Human Perception and Performance* **19** 580-597.

Gonzalez, R., G. Wu. 1999. On the form of the probability weighting function. *Cognitive Psychology* **38** 129-166.

Gul, F. 1991. A theory of disappointment aversion. *Econometrica* **59** 667-686.

Harless, D., C. F. Camerer. 1994. The predictive utility of generalized expected utility theories. *Econometrica* **62** 1251-1289.

Hershey, J. C., P. J. H. Schoemaker. 1985. Probability versus certainty equivalence methods in utility measurement: Are they equivalent? *Management Science* **31** 1213-1231.

Hertwig, R., A. Ortmann. 2001. Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences* **24** 383-451.

Kahneman, D., A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* **47** 263-291.

Keller, R. L., D. N. Kleinmuntz. 1998. Is this the right time for a new decision analysis journal? *Decision Analysis Society Newsletter* **17**

Llewellyn-Thomas, H., H. J. Sutherland, R. Tibshirani, A. Ciampi, J. E. Till, N. F. Boyd. 1982. The measurement of patients' values in medicine. *Medical Decision Making* **2** 449-462.

McCord, M., R. de Neufville. 1986. Lottery equivalents: Reduction of the certainty effect problem in utility assessment. *Management Science* **32** 56-60.

Miyamoto, J. M., S. A. Eraker. 1985. Parameter estimates for a qaly utility model. *Medical Decision Making* **5** 191-213.

Miyamoto, J. M., S. A. Eraker. 1988. A multiplicative model of the utility of survival duration and health quality. *Journal of Experimental Psychology: General* **117** 3-20.

Morrison, G. C. 2000. The endowment effect and expected utility. *Scottish Journal of Political Economy* **47** 183-197.

Oliver, A. 2003. The internal consistency of the standard gamble: Tests after adjusting for prospect theory. *Journal of Health Economics* **22** 659-674.

Payne, J. W., J. R. Bettman, D. A. Schkade. 1999. Measuring constructed preferences: Towards a building code. *Journal of Risk and Uncertainty* **19** 243-270.

Payne, J. W., D. J. Laughhunn, R. Crum. 1980. Translation of gambles and aspiration level effects in risky choice behavior. *Management Science* **26** 1039-1060.

Payne, J. W., D. J. Laughhunn, R. Crum. 1981. Further tests of aspiration level effects in risky choice behavior. *Management Science* **27** 953-958.

Pinto, J. L., J. M. Abellan-Perpiñan. 2005. Measuring the health of populations: The veil of ignorance approach. *Health Economics* **14** 69-82.

Pliskin, J. S., D. S. Shepard, M. C. Weinstein. 1980. Utility functions for life years and health status. *Operations Research* **28** 206-223.

Quiggin, J. 1981. Risk perception and risk aversion among australian farmers. *Australian Journal of Agricultural Economics* **25** 160-169.

Robinson, A., G. Loomes, M. Jones-Lee. 2001. Visual analog scales, standard gambles, and relative risk aversion. *Medical Decision Making* **21** 17-27.

Rottenstreich, Y., C. K. Hsee. 2001. Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science* **12** 185-190.

Rutten-van Mölken, M. P., C. H. Bakker, E. K. A. van Doorslaer, S. van der Linden. 1995. Methodological issues of patient utility measurement. Experience from two clinical trials. *Medical Care* **33** 922-937.

Shalev, J. 2000. Loss aversion equilibrium. *International Journal of Game Theory* **29** 269-287.

Smith, J. E., D. von Winterfeldt. 2004. Decision analysis in *management science*. *Management Science* **50** 561-574.

Stalmeier, P. F. M., T. G. G. Bezembinder. 1999. The discrepancy between risky and riskless utilities: A matter of framing? *Medical Decision Making* **19** 435-447.

Starmer, C. 2000. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* **28** 332-382.

Stiggelbout, A. M., G. M. Kiebert, J. Kievit, J. W. H. Leer, G. Stoter, J. C. J. M. de Haes. 1994. Utility assessment in cancer patients: Adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. *Medical Decision Making* **14** 82-90.

Sugden, R. 2003. Reference-dependent subjective expected utility. *Journal of Economic Theory* **111** 172-191.

Tversky, A., C. Fox. 1995. Weighing risk and uncertainty. *Psychological Review* **102** 269-283.

Tversky, A., D. Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* **5** 297-323.

Tversky, A., S. Sattath, P. Slovic. 1988. Contingent weighting in judgment and choice. *Psychological Review* **95** 371-384.

Wu, G., R. Gonzalez. 1996. Curvature of the probability weighting function. *Management Science* **42** 1676-1690.